Risk-Averse No-Regret Learning in Online Convex Games

Zifan Wang [1] **Yi Shen** [2] **Michael M. Zavlanos** [2]

Abstract

We consider an online stochastic game with risk-averse agents whose goal is to learn optimal decisions that minimize the risk of incurring significantly high costs. Specifically, we use the Conditional Value at Risk (CVaR) as a risk measure that the agents can estimate using bandit feedback in the form of the cost values of only their selected actions. Since the distributions of the cost functions depend on the actions of all agents that are generally unobservable, they are themselves unknown and, therefore, the CVaR values of the costs are difficult to compute. To address this challenge, we propose a new online risk-averse learning algorithm that relies on one-point zeroth-order estimation of the CVaR gradients computed using CVaR values that are estimated by appropriately sampling the cost functions. We show that this algorithm achieves sub-linear regret with high probability. We also propose two variants of this algorithm that improve performance. The first variant relies on a new sampling strategy that uses samples from the previous iteration to improve the estimation accuracy of the CVaR values. The second variant employs residual feedback that uses CVaR values from the previous iteration to reduce the variance of the CVaR gradient estimates. We theoretically analyze the convergence properties of these variants and illustrate their performance on an online market problem that we model as a Cournot game.

1. Introduction

Online convex optimization (OCO) aims at solving optimization problems with unknown cost functions using only samples of the cost function values. Many practical applications can be modeled as OCO problems. Examples include spam filtering (Hazan, 2019) and portfolio management (Hazan, 2006), among many others (Shalev-Shwartz et al., 2011). Oftentimes, OCO problems involve multiple agents interacting with each other in the same environment; for instance, in traffic routing (Sessa et al., 2019) and economic market optimization (Shi & Zhang, 2019), agents cooperate or compete, respectively, by sequentially selecting the best decisions that minimize their expected accumulated costs. These problems can be formulated as online convex games (Shalev-Shwartz & Singer, 2006; Gordon et al., 2008), and constitute the focus of this paper.

Typically, the performance of online optimization algorithms is measured using different notions of regret (Hazan, 2019), that capture the difference between the agents' online decisions and the optimal decisions in hindsight. An online algorithm is said to be no-regret (no-external-regret) if its regret is sub-linear in time (Gordon et al., 2008), i.e., if the agents are able to eventually learn the optimal decisions. Many no-regret algorithms have been proposed and analyzed for online convex games including (Shalev-Shwartz & Singer, 2006; Gordon et al., 2008; Hazan, 2019; Shalev-Shwartz et al., 2011). Common in these problems is the objective of the agents to minimize their expected cost functions. However, in high-stakes applications, minimizing the expected cost alone is not sufficient; avoiding the worst case is equally important. For example, in portfolio management, investing in the assets that yield the highest expected return rate is not necessarily the best decision since these assets may also be highly volatile and result in severe losses. To control for such catastrophic events, appropriate risk-averse criteria need to be considered during optimization, such as the Sharpe Ratio (Sharpe, 1994) or Conditional Value at Risk (CVaR) (Artzner et al., 1999).

In this paper, we consider online convex games with risk-averse agents, whose goal is to minimize the CVaR values of their cost functions. Moreover, we assume that only bandit feedback in the form of the costs of selected actions is available to the agents to estimate their CVaR values. To the best of our knowledge, risk-averse learning in convex games has not been explored in the literature. Most closely related to the problem considered here is work on deterministic online convex games with bandit feedback, including

[1]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. [2]Department of Mechanical Engineering & Material Science, Duke University, Durham, NC 27708, USA. Correspondence to: Zifan Wang <zifanwang199710@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

(Bravo et al., 2018; Duvocelle et al., 2018; Tatarenko & Kamgarpour, 2018; Lin et al., 2020). Specifically, (Bravo et al., 2018) relies on tools from stochastic approximation theory to show that derivative-free methods for monotone and concave games converge to the Nash equilibrium with probability 1. This work is extended in (Duvocelle et al., 2018) to time-varying games. The authors in (Duvocelle et al., 2018) show that if the time-varying game converges, then the sequence of actions converges to the Nash equilibrium. Common in these works is that the cost functions are deterministic. As such, they cannot model risk in the presence of uncertainty. Methods for risk-averse learning have been investigated, e.g., in (Urpí et al., 2021; Kalogerias & Powell, 2019; Chow et al., 2017). Specifically, in (Urpí et al., 2021), a risk-averse offline reinforcement learning algorithm is proposed that exhibits better performance compared to risk-neural approaches for robot control tasks. In (Kalogerias & Powell, 2019), a zeroth-order method for mean-semideviation-based risk-averse learning is proposed. We note that, despite the importance of controlling risk in many applications, only a few works employ CVaR as a risk measure and still provide theoretical results, e.g., (Curi et al., 2019; Cardoso & Xu, 2019; Tamkin et al., 2019; Soma & Yoshida, 2020; Kalogerias, 2020). In (Curi et al., 2019), risk-averse learning is transformed into a zero-sum game between a sampler and a learner. Then, using an adaptive sampling strategy, the regret of this game is analyzed. In (Tamkin et al., 2019), a sub-linear regret algorithm is proposed for risk-averse multi-arm bandit problems by constructing empirical cumulative distribution functions for each arm from online samples. Recently, (Kalogerias, 2020) has shown that CVaR learning problems subject to not necessarily convex loss functions can be solved as efficiently as their risk-neutral counterparts.

Compared to the literature discussed above, risk-averse learning for online convex games possesses unique challenges, including: (1) The distribution of an agent's cost function depends on other agents' actions, and (2) Using finite bandit feedback, it is difficult to accurately estimate the continuous distributions of the cost functions and, therefore, accurately estimate the CVaR values. To address these challenges, in this paper we use samples of the cost functions to learn an empirical distribution function (EDF) of the random costs. Then, using this EDF, the agents can estimate the CVaR values of their cost functions, and use these CVaR values to construct zeroth-order estimates of the CVaR gradients. By appropriately designing this sampling strategy, we show that with high probability, the accumulated error of the CVaR estimates is bounded, and the accumulated error of the zeroth-order CVaR gradient estimates is also bounded. As a result, our method achieves sub-linear regret with high probability. To further improve the regret of our method, we allow our sampling strategy to use previous samples to reduce the accumulated error of the CVaR estimates. Specifically, at time step t, we build the EDF estimate using samples from times t and $t-1$, and then use this EDF to estimate the CVaR values and the corresponding CVaR gradients, as before. Assuming that the variation of the CDF of the cost function at two consecutive time steps is bounded by the distance between the two corresponding actions at these time steps, we theoretically show that the accumulated error of the CVaR estimates is strictly less than that achieved without reusing previous samples under certain conditions. We also provide an alternative way of improving the regret by utilizing residual feedback (Zhang et al., 2020a;b) that reduces the variance of the zeroth-order CVaR gradient estimates. We illustrate our method on an online market problem that we model as a Cournot game (Allaz & Vila, 1993).

To the best of our knowledge, this is the first work to address risk-averse learning in online convex games. Note that the CVaR value of each agent depends on the joint actions of all agents, hence the proposed risk-averse game is in essence a time-varying game as the agents update their actions sequentially. All existing literature on learning in games discussed before considers static games, except for (Duvocelle et al., 2018), that requires knowledge of whether the game converges and how the Nash equilibrium changes. However, the time-varying nature of the game considered here is due to the updates of the other agents and, therefore, it is not possible to know *a prior* whether this game will converge or not. As a result, the analysis in (Duvocelle et al., 2018) cannot be applied to analyze the game considered here. In addition, existing literature that employs zeroth-order techniques to solve learning problems in games typically relies on constructing unbiased gradient estimates of the smoothed cost functions. Nevertheless, unbiased gradient estimates cannot be obtained in risk-averse games since it is not possible to obtain accurate CVaR estimates of the cost functions merely using finite bandit feedback. Perhaps closest to the method proposed here is the approach in (Cardoso & Xu, 2019), that makes a first attempt to analyze risk-averse bandit learning problems. Using CVaR properties, the authors reformulate the CVaR optimization problem to an equivalent optimization problem of an augmented L function and show that the reformulated problem converges in the single-agent case. However, the analysis in (Cardoso & Xu, 2019) cannot be easily extended to multi-agent problems since minimizing the L functions is not equivalent to minimizing CVaR values in multi-agent games.

The rest of the paper is organized as follows. In Section 2, we define the proposed risk-averse online game and provide some assumptions. The main algorithm is presented in Section 3 with corresponding regret analysis. Two variants of this algorithm with improved regrets are provided in Section 4. In section 5, we use an online market example

to illustrate the effectiveness of the proposed algorithms. Finally, we conclude this work in Section 6.

2. Problem Definition

We consider a repeated game $\mathcal{G}$ with N agents. At the beginning of each episode, each agent $i \in \mathcal{N} = \{1, \dots, N\}$ simultaneously chooses an action $x_i \in \mathbb{R}^{d_i}$ from a convex set $\mathcal{X}_i$ and receives a random cost value that is sampled from the cost function $J_i(x_i, x_{-i}, \xi_i) : \mathcal{X} \times \Xi_i \to \mathbb{R}$, where x_i is the action of agent i, x_{-i} denotes the actions of agents except for agent i, $\mathcal{X} = \Pi_{i=1}^N \mathcal{X}_i$ is the joint action space and $\xi_i \in \Xi_i$ describes the uncertainty of the cost function. Here we assume that the diameter of the convex set $\mathcal{X}_i$ is bounded by D_x for all $i = 1, \dots, N$. For ease of notation, we sometimes denote the cost function as $J_i(x, \xi_i)$, where $x = (x_i, x_{-i})$ is the concatenated vector of all agents' actions.

We use the Conditional Value at Risk (CVaR) as a risk measure to model the risk-aversion in the agents. Specifically, suppose that the random variable $J_i(x, \xi_i)$ has the CDF $F_x(y) = \mathbb{P}\{J_i(x, \xi_i) \leq y\}$. We drop the decision variable x in F_x whenever it is clear from the contexts. Then, for a given risk level $\alpha_i \in [0, 1]$, the CVaR of the cost function $J_i(x, \xi_i)$ of agent i at point x is defined as

$$\begin{aligned} C_i(x) :&= \mathrm{CVaR}_{\alpha_i}[J_i(x, \xi_i)] \\ &= \mathbb{E}_F[J_i(x, \xi_i) | J_i(x, \xi_i) \geq J^{\alpha_i}], \end{aligned}$$

where J^{α_i} is the $1 - \alpha_i$ quantile of the distribution, which is also termed as Value at Risk (VaR). CVaR captures the average cost under the tail of the distribution of $J_i(x, \xi_i)$. Note that the CVaR value of the random variable $J_i(x, \xi_i)$ is determined its cumulative distribution function F, hence we sometimes write CVaR as a function of the CDF, i.e., $\mathrm{CVaR}_{\alpha_i}[J_i(x, \xi_i)] = \mathrm{CVaR}_{\alpha_i}[F]$ for ease of notation.

Here we assume that the agents have no prior knowledge about this game, i.e, the agents do not know the cost functions $J_i(x, \xi_i)$, and cannot observe the other agents' actions. The only information that is available to the agents is the cost of the selected actions. Moreover, we assume that the cost function $J_i(x, \xi_i)$ of each agent satisfies the following assumptions.

Assumption 1. *The function $J_i(x_i, x_{-i}, \xi_i)$ is convex in x_i for every $\xi_i \in \Xi_i$ and bounded by U, i.e., $|J_i(x, \xi_i)| \leq U$, for all $i = 1, \dots, N$.*

Assumption 2. *$J_i(x, \xi_i)$ is L_0-Lipschitz continuous in x for every $\xi_i \in \Xi_i$, for all $i = 1, \dots, N$.*

Assumptions 1 and 2 hold in many applications, e.g., the Cournot game (Shi & Zhang, 2019) and the repeated Kelly auctions (Duvocelle et al., 2018).

Given Assumptions 1 and 2, the below well known result characterizes important properties of the CVaR function. The proof can be found in (Cardoso & Xu, 2019).

Lemma 1. *Given Assumptions 1 and 2, we have that $C_i(x_i, x_{-i})$ is convex in x_i and L_0-Lipschitz continuous in x, for all $i = 1, \dots, N$.*

The objective of each agent i is to minimize its cumulative CVaR functions. Specifically, given a sequence of agents' actions $\{\hat{x}_t\}_{t=1}^T$ over T episodes, where $\hat{x}_t = (\hat{x}_{i,t}, \hat{x}_{-i,t})$ denotes the agents' actual played actions at time step t, we define the regret (CVaR-regret) for agent i as

$$\mathrm{R}_{C_i}(T) = \sum_{t=1}^{T} C_i(\hat{x}_{i,t}, \hat{x}_{-i,t}) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i(\tilde{x}_i, \hat{x}_{-i,t}),$$

which measures the cumulative loss against a best single policy in hindsight. Then, our goal in this paper is to design a no-regret (equivalently, sub-linear regret) algorithm to solve this game, such that $\lim_{T \to \infty} \frac{\mathrm{R}_{C_i}(T)}{T} = 0$ for all agents.

3. A Risk-Averse Learning Algorithm

In this section, we propose a risk-averse learning algorithm to solve the proposed online convex game. Our algorithm relies on a novel sampling strategy to estimate the CVaR values and a one-point zeroth-order estimator of the CVaR gradient. Specifically, since estimation of CVaR values requires the distribution of the cost functions which is impossible to compute using a single evaluation of the cost functions per time step, we assume that the agents can sample the cost functions multiple times to learn their distributions. For this, we introduce a practical sampling strategy described below.

During each time step t, the agents keep their actions fixed and draw n_t samples of their individual cost functions. Then, they use these samples to determine their actions for time step $t + 1$ and then sample again. The sampling strategy is defined as

$$n_t = \lceil bU^2(T - t + 1)^a \rceil, \tag{1}$$

where $\lceil \cdot \rceil$ is the ceiling function, T is the time horizon, U is the cost function bound as in Assumption 1, and $a, b \in (0, 1)$ are parameters to be selected later. The parameters a, b are assumed to be known by all the agents beforehand so that the game is synchronous. Moreover, since $a < 1$, the number of samples n_t will decrease with the iterations and eventually equal to 1.

The proposed risk-averse learning algorithm for online convex games is illustrated in Algorithm 1. At time step t, the agents randomly perturb their current actions $x_{i,t}$ by an amount $\delta u_{i,t}$, where $u_{i,t} \in \mathbb{S}^{d_i}$ is a random perturbation direction sampled from a unit sphere $\mathbb{S}^{d_i} \subset \mathbb{R}^{d_i}$ and δ is

Algorithm 1 Risk-averse learning

Require: Initial value x_0, step size η, parameters a, b, δ, T, risk level α_i, $i = 1, \cdots, N$.

for *episode* $t = 1, \ldots, T$ **do**
  Select $n_t = \lceil bU^2(T-t+1)^a \rceil$
  Each agent samples $u_{i,t} \in \mathbb{S}^{d_i}$, $i = 1, \ldots, N$
  Each agent play $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$, $i = 1, \ldots, N$
  for $j = 1, \ldots, n_t$ **do**
    Let all agents play $\hat{x}_{i,t}$
    Obtain $J_i(\hat{x}_{i,t}, \hat{x}_{-i,t}, \xi_i^j)$
  end for
  for agent $i = 1, \ldots, N$ **do**
    Build EDF $\hat{F}_{i,t}(y)$
    Calculate CVaR estimate: $\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]$
    Construct gradient estimate $\hat{g}_{i,t} = \frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]u_{i,t}$
    Update x: $x_{i,t+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta\hat{g}_{i,t})$
  end for
end for

the size of this perturbation. Then, using the sampling strategy defined above, the agents play their perturbed actions $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$ for n_t times, and obtain n_t samples of their cost functions. For agent i, at time step t, we denote the CDF of the random cost $J_i(\hat{x}_t, \xi_i)$ that is returned by the perturbed action $\hat{x}_t$ as $F_{i,t}(y) = \mathbb{P}\{J_i(\hat{x}_t, \xi_i) \leq y\}$. Since the agents cannot accurately estimate this continuous CDF $F_{i,t}(y)$ using finite samples, they instead construct the EDF of $J_i(\hat{x}_t, \xi_i)$ as

$$\hat{F}_{i,t}(y) = \frac{1}{n_t}\sum_{j=1}^{n_t} \mathbf{1}\{J_i(\hat{x}_t, \xi_i^j) \leq y\}, \tag{2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Then, using this EDF, the agents construct CVaR estimates of their cost functions $J_i(\hat{x}_t, \xi_i)$, denoted as $\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]$, which they use to further construct zeroth-order estimates of their CVaR gradients as

$$\hat{g}_{i,t} = \frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]u_{i,t}, \tag{3}$$

where δ is the size of the perturbation on the action $x_{i,t}$ defined above. To ensure that the function values at the queried points during each time step are always feasible, we define the projection set $\mathcal{X}_i^\delta = \{x_i \in \mathcal{X}_i | \mathrm{dist}(x_i, \partial\mathcal{X}_i) \geq \delta\}$. Then, the agents perform the following projected gradient-descent update

$$x_{i,t+1} = \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta\hat{g}_{i,t}). \tag{4}$$

To analyze Algorithm 1, we utilize the smoothed approximation of $C_i(x)$ defined as $C_i^\delta(x) = \mathbb{E}_{w_i \sim \mathbb{B}_i, u_{-i} \sim \mathbb{S}_{-i}}[C_i(x_i + \delta w_i, x_{-i} + \delta u_{-i})]$, where $\mathbb{S}_{-i} = \Pi_{j \neq i}\mathbb{S}_j$, and $\mathbb{B}_i, \mathbb{S}_i$ denote the unit ball and unit sphere in $\mathbb{R}^{d_i}$, respectively, and the size of the perturbation δ here serves as a smoothing parameter that controls how well $C_i^\delta(x)$ approximates $C_i(x)$. For details on zeroth-order optimization methods and their analysis, see (Nesterov & Spokoiny, 2017). The function $C_i^\delta(x)$ satisfies the following properties. The proof can be found in Appendix A.1.

Lemma 2. *Let Assumptions 1 and 2 hold. Then we have that*

1. *$C_i^\delta(x_i, x_{-i})$ is convex in x_i,*
2. *$C_i^\delta(x)$ is L_0-Lipschitz continuous in x,*
3. *$|C_i^\delta(x) - C_i(x)| \leq \delta L_0\sqrt{N}$.*

From Lemma C.1 in (Bravo et al., 2018) , we have that

$$\mathbb{E}[\frac{d_i}{\delta}C_i(\hat{x}_t)u_{i,t}] = \nabla_i C_i^\delta(x_t), \tag{5}$$

where ∇_i denotes the partial derivative with respect to x_i. However, as discussed before, it is not possible to accurately estimate the CVaR value $C_i(\hat{x}_t)$ using finite samples of the cost function $J_i(\hat{x}_t, \xi_i)$. Instead, there will exist a CVaR estimation error, which we define as

$$\hat{\varepsilon}_{i,t} := \mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t}].$$

Then, we have that $\mathbb{E}[\hat{g}_{i,t}] = \mathbb{E}\left[\frac{d_i}{\delta}(C_i(\hat{x}_t) + \hat{\varepsilon}_{i,t})u_{i,t}\right] = \nabla_i C_i^\delta(x_t) + \mathbb{E}\left[\frac{d_i}{\delta}\hat{\varepsilon}_{i,t}u_{i,t}\right]$, which indicates that the CVaR gradient estimate is biased due to the use of finite samples. The analysis of the CVaR estimation error $\hat{\varepsilon}_{i,t}$ plays a key role in the whole analysis of the regret of Algorithm 1. To bound the CVaR estimation error, we first present a lemma that bounds the difference between the CVaR values for two different CDFs. The proof of this lemma can be found in Appendix A.2.

Lemma 3. *Let F and G be two CDFs of two random variables and the random variables are bounded by U. Then we have that*

$$|\mathrm{CVaR}_\alpha[F] - \mathrm{CVaR}_\alpha[G]| \leq \frac{U}{\alpha}\sup_y |F(y) - G(y)|.$$

Lemma 3 states that the distance between two CVaR values is related to the distance between the corresponding CDFs. By substituting $F = F_{i,t}$ and $G = \hat{F}_{i,t}$ into Lemma 3 and applying the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, we have that

$$\begin{aligned} |\hat{\varepsilon}_{i,t}| &= |\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t}]| \\ &\leq \frac{U}{\alpha_i}\sqrt{\frac{\ln(2/\bar{\gamma})}{2n_t}} \end{aligned} \tag{6}$$

with probability at least $1-\bar{\gamma}$. Combining inequality (6) with the sampling strategy defined in equation (1), the accumulated error of CVaR estimation can be bounded, which is given in the following lemma whose proof can be found in Appendix A.3.

Lemma 4. *Given a confidence level $\bar{\gamma}$ and the sampling strategy in equation (1), we have that the following inequality holds:*

$$\sum_{t=1}^{T} |\hat{\varepsilon}_{i,t}| \leq B_1 \tag{7}$$

with probability at least $1-\gamma$, where $\gamma = \bar{\gamma}T$ and $B_1 = \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{b}}T^{1-\frac{a}{2}}$.

The following result provides a generic regret decomposition of Algorithm 1. The proof can be found in Appendix A.4.

Lemma 5. *Let Assumptions 1 and 2 hold. Then, the regret of Algorithm 1 satisfies*

$$\mathrm{R}^1_{C_i}(T) \leq Err(ZO) + Err(CVaR),$$

where $Err(ZO) = \frac{D_x^2}{2\eta} + \frac{d_i^2 U^2 \eta}{2\delta^2}T + (4\sqrt{N}+\Omega)L_0\delta T$, $Err(CVaR) = \frac{d_i D_x}{\delta}B_1$, $\Omega > 0$ is a constant that represents the error from projection $\mathcal{P}$, and B_1 as in (7).

Lemma 5 decomposes the regret into two terms, a zeroth-order error term and a CVaR estimation error term. By selecting η and δ appropriately, we can show that Algorithm 1 is no-regret. In the following theorem, $\tilde{\mathcal{O}}$ hides constant factors and poly-logarithmic factors of T. In contrast, the standard notation $\mathcal{O}$ only hides constant factors.

Theorem 1. *Let Assumptions 1 and 2 hold and select $\delta = \frac{\sqrt{D_x U d_i}}{N^{\frac{1}{4}}T^{\frac{a}{4}}\sqrt{\alpha_i L_0}}$, $\eta = \frac{\sqrt{\alpha_i}D_x^{\frac{3}{2}}}{\sqrt{L_0 U d_i}N^{\frac{1}{4}}T^{\frac{3a}{4}}}$. Suppose that n_t is chosen as in equation (1) with $a \in (0,1)$, and the EDF and the gradient estimate are defined as in equations (2) and (3), respectively. Then, Algorithm 1 achieves regret $\mathrm{R}^1_{C_i}(T) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$ with probability at least $1-\gamma$.*

Proof. Substituting δ, η and B_1 into the regret bound $Err(\mathrm{ZO}) + Err(\mathrm{CVaR})$ in Lemma 5, we obtain that $\mathrm{R}^1_{C_i}(T) \leq Err(\mathrm{ZO}) + Err(\mathrm{CVaR}) = \mathcal{O}(\sqrt{D_x U d_i L_0}N^{\frac{1}{4}}\alpha_i^{-\frac{1}{2}}\sqrt{\ln(T/\gamma)}T^{1-\frac{a}{4}}) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$. The proof is complete. □

As shown in Theorem 1, although it is impossible to obtain accurate CVaR values using finite bandit feedback, our method still achieves sub-linear regret with high probability. Notice that the choice of the risk level α_i can also affect the regret. Specifically, a lower value for α_i can result in higher regret, since it is harder to get samples under the α_i tail of the distribution.

4. Improving the Algorithm Regret

In this section, we propose two variants of Algorithm 1 that improve the regret. The first variant reduces the accumulated error of the CVaR estimates by using samples from the previous iteration. The second variant employs residual feedback (Zhang et al., 2020a;b) to reduce the variance of the CVaR gradient estimates. A relevant analysis is given in the following two subsections, respectively.

4.1. Improving the CVaR Estimation Accuracy

The accuracy of the CVaR estimation in Algorithm 1 depends on the number of samples of the cost functions at each iteration according to equation (6); the more samples, the better the CVaR estimation accuracy. To further improve the CVaR estimation accuracy, we propose a modification to Algorithm 1 that reuses samples from the previous iteration, effectively increasing the number of available samples per iteration while maintaining the number of new samples the same. First, we make the following assumption on the variation of the cumulative distribution function.

Assumption 3. *Let $F_{i,t}(y) = \mathbb{P}\{J_i(\hat{x}_t, \xi_i) \leq y\}$ and $F_{i,t-1}(y) = \mathbb{P}\{J_i(\hat{x}_{t-1}, \xi_i) \leq y\}$. There exist constants $C_1, C_2 > 0$ such that*

$$\sup_y |F_{i,t}(y) - F_{i,t-1}(y)| \leq (C_1\delta + C_2)\|x_t - x_{t-1}\|.$$

Assumption 3 states that the variation of the CDF across two consecutive time steps is bounded by the distance between the corresponding unperturbed actions. It means that if x_t is close to x_{t-1}, then the corresponding cost function values $J_i(\hat{x}_t, \xi_i)$ and $J_i(\hat{x}_{t-1}, \xi_i)$ for every ξ_i should also be close to each other, and so should be the two CDFs. Note that the bound in Assumption 3 is also related to the smoothing parameter δ, since the played action is in fact the perturbed one, i.e., $\hat{x}_t$. Moreover, note that this bound cannot go to 0 by decreasing the smoothing parameter δ, which implies that the variation in the CDF should be dominated by the distance between x_t and x_{t-1}.

The proposed risk-averse learning algorithm with sample reuse is illustrated in Algorithm 2 that can be found in Appendix B. Specifically, assuming that the agents can sample the cost functions n_t times at every time step t, for $t \geq 2$, we define a new EDF as

$$\tilde{F}_{i,t}(y) = \frac{n_t}{N_t}\hat{F}_{i,t} + \frac{n_{t-1}}{N_t}\hat{F}_{i,t-1}, \tag{8}$$

where $N_t = n_t + n_{t-1}$. For $t = 1$, we set the initial value as $\tilde{F}_{i,1} = \hat{F}_{i,1}$ and $N_1 = n_1$. Using this sampling strategy, we design the CVaR gradient estimate as

$$\tilde{g}_{i,t} = \frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\tilde{F}_{i,t}]u_{i,t}. \tag{9}$$

Note that, as in Algorithm 1, this gradient estimate is biased since the estimation of CVaR uses not only finite samples, but also samples from the previous iteration. We denote the CVaR estimation error as

$$\tilde{\varepsilon}_{i,t} := \mathrm{CVaR}_{\alpha_i}[\tilde{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t}]. \tag{10}$$

Due to the use of previous samples, the analysis of the CVaR estimation error in this case becomes more complicated. The following lemma characterizes the CVaR estimation error and its the proof can be found in Appendix B.1.

Lemma 6. *Given a confidence level γ, the following inequality holds*

$$\begin{aligned}|\tilde{\varepsilon}_{i,t}| \leq & \frac{U}{\alpha_i}\left(\sqrt{\frac{\ln(2T/\gamma)}{2(n_t+n_{t-1})}}\right) \\ & + \frac{U}{\alpha_i}\left(\frac{(C_1\delta+C_2)d_iU\sqrt{N}\eta}{2\delta}\right),\end{aligned} \tag{11}$$

with probability at least $1-\gamma$, for $\forall t = 2,\dots,T$.

Using the above bound on the CVaR estimation error $\tilde{\varepsilon}_{i,t}$, we are able to show the following result.

Lemma 7. *Assume the same values for δ and η as in Algorithm 1, i.e., $\delta = \frac{\sqrt{D_xUd_i}}{N^{\frac{1}{4}}\sqrt{\alpha_iL_0}}T^{-\frac{a}{4}}$, and $\eta = \frac{\sqrt{\alpha_i}D_x^{\frac{3}{2}}}{\sqrt{L_0Ud_i}N^{\frac{1}{4}}}T^{-\frac{3a}{4}}$. Then, given any constant $\lambda > 0$, there exists $T_\lambda > 0$ such that when $T > T_\lambda$, we have $\sum_{t=1}^T |\tilde{\varepsilon}_{i,t}| \leq B_1 - \lambda T^{1-\frac{3a}{4}}$, with probability at least $1-\gamma$.*

Proof. For ease of notation, we let $\delta = \Sigma_1 T^{-\frac{a}{4}}$ and $\eta = \Sigma_2 T^{-\frac{3a}{4}}$, where $\Sigma_1 = \frac{\sqrt{D_xUd_i}}{N^{\frac{1}{4}}\sqrt{\alpha_iL_0}}$, $\Sigma_2 = \frac{\sqrt{\alpha_i}D_x^{\frac{3}{2}}}{\sqrt{L_0Ud_i}N^{\frac{1}{4}}}$. Summing equation (11) over t, we obtain

$$\begin{aligned}&\sum_{t=2}^T |\tilde{\varepsilon}_{i,t}| \\ &\leq \frac{U}{\alpha_i}\sum_{t=2}^T\left(\sqrt{\frac{\ln(2T/\gamma)}{2(n_t+n_{t-1})}} + \frac{(C_1\delta+C_2)d_iU\sqrt{N}\eta}{2\delta}\right) \\ &\leq \frac{U}{\alpha_i}\sum_{t=2}^T\sqrt{\frac{\ln(2T/\gamma)}{4n_{t-1}}} + \frac{(C_1\delta+C_2)d_iU^2\sqrt{N}\eta T}{2\alpha_i\delta} \\ &\leq \sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}T^{1-\frac{a}{2}} + \frac{C_1d_iU^2\sqrt{N}\Sigma_2}{2\alpha_i}T^{1-\frac{3a}{4}} \\ &\quad + \frac{C_2d_iU^2\sqrt{N}\Sigma_2}{2\alpha_i\Sigma_1}T^{1-\frac{a}{2}}.\end{aligned}$$

Adding the first term $|\tilde{\varepsilon}_{i,1}|$ to both sides of this inequality, we have that

$$\begin{aligned}&\sum_{t=1}^T |\tilde{\varepsilon}_{i,t}| - B_1 \\ &\leq -(\sqrt{2}-1)\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}T^{1-\frac{a}{2}} + \Sigma_3 T^{1-\frac{3a}{4}} \\ &\quad + \Sigma_4 T^{1-\frac{a}{2}} + |\tilde{\varepsilon}_{i,1}| \\ &\leq \left(-(\sqrt{2}-1)\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}} + \Sigma_4\right)T^{1-\frac{a}{2}} \\ &\quad + \left(\Sigma_3 + |\tilde{\varepsilon}_{i,1}|\right)T^{1-\frac{3a}{4}} \\ &\leq f(T)T^{1-\frac{3a}{4}},\end{aligned} \tag{12}$$

where $\Sigma_3 = \frac{C_1d_iU^2\sqrt{N}\Sigma_2}{2\alpha_i}$ and $\Sigma_4 = \frac{C_2d_iU^2\sqrt{N}\Sigma_2}{2\alpha_i\Sigma_1}$ and $f(T) := \Sigma_3 + |\tilde{\varepsilon}_{i,1}| - \left((\sqrt{2}-1)\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}} - \Sigma_4\right)T^{\frac{a}{4}}$. Observe that the function $f(T)$ is monotonically decreasing in T and approaches negative infinity when $T \to \infty$. Hence there exists T_λ with $f(T_\lambda) = -\lambda$ such that when $T > T_\lambda$ we have $f(T) < -\lambda$, and thus $\sum_{t=1}^T |\tilde{\varepsilon}_{i,t}| - B_1 \leq -\lambda T^{1-\frac{3a}{4}}$. The proof is complete. □

This lemma shows that for the same confidence level $1-\gamma$, selecting $\delta = \Sigma_1 T^{-\frac{a}{4}}$ and $\eta = \Sigma_2 T^{-\frac{3a}{4}}$ in Algorithm 2 results in a bound on the accumulated CVaR estimation error that is strictly less than that achieved by Algorithm 1. The proof can be found in Appendix B.2. Similar to Lemma 5, we can decompose the regret into two sources of errors as shown below.

Lemma 8. *Let Assumptions 1 and 2 hold. Then, the regret of Algorithm 2 satisfies*

$$\mathrm{R}^2_{C_i}(T) \leq Err(ZO) + \frac{d_iD_x}{\delta}\sum_{t=1}^T |\tilde{\varepsilon}_{i,t}|, \tag{13}$$

where $Err(ZO)$ is the zeroth order error term as in Lemma 5.

Recall that the regret achieved by Algorithm 1 is bounded by $Err(\mathrm{ZO}) + Err(\mathrm{CVaR}) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$. In what follows, we show that the regret achieved by Algorithm 2 is strictly smaller than the regret bound achieved by Algorithm 1, i.e., $Err(\mathrm{ZO}) + Err(\mathrm{CVaR})$.

Theorem 2. *Let Assumptions 1, 2 and 3 hold, and assume the same values for δ and η as in Algorithm 1, i.e., $\delta = \frac{\sqrt{D_xUd_i}}{N^{\frac{1}{4}}\sqrt{\alpha_iL_0}}T^{-\frac{a}{4}}$ and $\eta = \frac{\sqrt{\alpha_i}D_x^{\frac{3}{2}}}{\sqrt{L_0Ud_i}N^{\frac{1}{4}}}T^{-\frac{3a}{4}}$. Suppose that n_t is chosen according to equation (1) with $a \in (0,1)$, and the EDF and the gradient estimate are defined as in equations (8) and (9), respectively. Then, when $T > T_\lambda$ with T_λ as in Lemma 7, Algorithm 2 achieves regret $\mathrm{R}^2_{C_i}(T) < Err(ZO) + Err(CVaR)$ and thus $\mathrm{R}^2_{C_i}(T) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$.*

Proof. Recall that $Err(\text{CVaR}) = \frac{d_i D_x}{\delta} B_1$. Adding and subtracting $Err(\text{CVaR})$ to the bound in Lemma 8, we have that

$$\begin{aligned}\text{R}^2_{C_i}(T) &\leq Err(\text{ZO}) + \frac{d_i D_x}{\delta} \sum_{t=1}^{T} |\tilde{\varepsilon}_{i,t}| \\ &= Err(\text{ZO}) + Err(\text{CVaR}) + \frac{d_i D_x}{\delta} (\sum_{t=1}^{T} |\tilde{\varepsilon}_{i,t}| - B_1).\end{aligned}$$

Combining this inequality with Lemma 7, and assuming that $T > T_\lambda$, we obtain that

$$\begin{aligned}\text{R}^2_{C_i}(T) &\leq Err(\text{ZO}) + Err(\text{CVaR}) \\ &= \mathcal{O}(\sqrt{D_x U d_i L_0} N^{\frac{1}{4}} \alpha_i^{-\frac{1}{2}} \sqrt{\ln(T/\gamma)} T^{1-\frac{a}{4}}),\end{aligned}$$

which completes the proof. □

Theorem 2 shows that when Assumption 3 holds, the regret bound achieved by using previous samples is guaranteed to be smaller than that achieved without using prior information.

Note that hybrid sampling strategies are also possible that initially use samples only from the current iteration and eventually switch to using samples from the previous iteration too. Assuming that t_0 denotes the switching time after which previous samples are reused, the EDF is defined as

$$\tilde{F}_{i,t}(y) = \begin{cases} \hat{F}_{i,t}, & t < t_0 + 1 \\ \frac{n_t}{N_t} \hat{F}_{i,t} + \frac{n_{t-1}}{N_t} \hat{F}_{i,t-1}, & t \geq t_0 + 1 \end{cases}, \quad (14)$$

Then, the regret achieved by this hybrid sampling strategy is analyzed below. The proof can be found in Appendix B.3.

Corollary 1. *Let Assumptions 1, 2 and 3 hold, and assume the same values for δ and η as in Algorithm 1, i.e., $\delta = \frac{\sqrt{D_x U d_i}}{N^{\frac{1}{4}} \sqrt{\alpha_i L_0}} T^{-\frac{a}{4}}$ and $\eta = \frac{\sqrt{\alpha_i} D_x^{\frac{3}{2}}}{\sqrt{L_0 U d_i} N^{\frac{1}{4}}} T^{-\frac{3a}{4}}$. Suppose that n_t is chosen according to equation (1) with $a \in (0,1)$, and the EDF and the gradient estimate are defined as in equations (14) and (9), respectively. Then, there exists $T_{\lambda(t_0)} > 0$ such that when $T > T_{\lambda(t_0)}$, Algorithm 2 achieves regret $\text{R}^2_{C_i}(T) < Err(ZO) + Err(CVaR)$ and thus $\text{R}^2_{C_i}(T) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$.*

4.2. Reducing the CVaR Gradient Estimation Variance

Algorithm 2 discussed in Section 4.1 reduces the CVaR estimation error by using samples from the previous iteration. In this section, we propose an alternative variation of Algorithm 1 that uses residual feedback (Zhang et al., 2020a;b) to reduce the variance of the zeroth-order CVaR gradient estimates. The risk-averse learning algorithm with residual feedback is illustrated in Algorithm 3 and can be found in Appendix C. Specifically, in Algorithm 3, the EDF is computed the same way as in Algorithm 1, but the gradient estimate now takes the form

$$\bar{g}_{i,t} = \frac{d_i}{\delta} (\text{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \text{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]) u_{i,t}. \quad (15)$$

The CVaR gradient estimation is still biased and we can define the error $\bar{\varepsilon}_{i,t} = \text{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \text{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}] - \text{CVaR}_{\alpha_i}[F_{i,t}]$. Recall the definition of $\hat{\varepsilon}_{i,t}$ in equation (10). Substituting in the expression for $\bar{\varepsilon}_{i,t}$, we have that $\bar{\varepsilon}_{i,t} = \hat{\varepsilon}_{i,t} - \text{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]$. Taking the expectation of the gradient in equation (15) with respect to $u_{i,t}$, we have that

$$\begin{aligned}\mathbb{E}[\bar{g}_{i,t}] &= \mathbb{E}[\frac{d_i}{\delta} (\bar{\varepsilon}_{i,t} + \text{CVaR}_{\alpha_i}[F_{i,t}]) u_{i,t}] \\ &= \nabla_i C_i^\delta(x_t) + \mathbb{E}\left[\frac{d_i}{\delta} \bar{\varepsilon}_{i,t} u_{i,t}\right] \\ &= \nabla_i C_i^\delta(x_t) + \mathbb{E}\left[\frac{d_i}{\delta} \left(\hat{\varepsilon}_{i,t} - \text{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]\right) u_{i,t}\right] \\ &= \nabla_i C_i^\delta(x_t) + \mathbb{E}\left[\frac{d_i}{\delta} \hat{\varepsilon}_{i,t} u_{i,t}\right], \quad (16)\end{aligned}$$

where the last equality is due to the fact that $u_{i,t}$ is independent of $\hat{F}_{i,t-1}$. Recall that the gradient estimate in Algorithm 1 satisfies $\mathbb{E}[\hat{g}_{i,t}] = \nabla_i C_i^\delta(x_t) + \mathbb{E}\left[\frac{d_i}{\delta} \hat{\varepsilon}_{i,t} u_{i,t}\right]$. Therefore, the expectation of the CVaR gradient estimate using residual feedback is the same as that in Algorithm 1. The following result analyzes the regret achieved by Algorithm 3. The proof can be found in Appendix C.

Theorem 3. *Let Assumptions 1 and 2 hold, and select $\eta = \frac{D_x}{d_i L_0 N} T^{-\frac{3a}{4}}$, $\delta = \frac{D_x}{N^{\frac{1}{6}}} T^{-\frac{a}{4}}$. Suppose that n_t is chosen according to equation (1) with $a \in (0,1)$, and the EDF and the gradient estimate are defined as in equations (2) and (15), respectively. Then, when $T \geq (8N^{\frac{2}{3}})^{\frac{1}{a}}$, Algorithm 3 achieves the regret $\text{R}^3_{C_i}(T) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$ with probability at least $1-\gamma$.*

More precisely, Algorithm 3 actually achieves the regret $\text{R}^3_{C_i}(T) = \mathcal{O}(D_x d_i L_0 N S(\alpha) \ln(T/\gamma) T^{1-\frac{a}{4}})$, where $S(\alpha) := \sum_{i=1}^{N} \frac{1}{\alpha_i^2}$; see Appendix C.2 for more details. Note that the poly-logarithmic term $\ln(T)$ in the regret bound achieved by Algorithm 3 is dominated by the polynomial term $T^{1-\frac{a}{4}}$ when T is large. Recall also that Algorithm 1 achieves regret $\text{R}^1_{C_i}(T) = \mathcal{O}(\sqrt{D_x U d_i L_0} N^{\frac{1}{4}} \alpha_i^{-\frac{1}{2}} \sqrt{\ln(T/\gamma)} T^{1-\frac{a}{4}})$. Ignoring the logarithmic dependence in the regret for large T, we obtain that the regret bound achieved by Algorithm 3 is strictly less than that achieved by Algorithm 1 when $U > D_x d_i L_0 N^{\frac{3}{2}} \alpha_i S^2(\alpha)$.

5. Numerical Experiments

Consider a Cournot game involving two risk-averse firms (agents) in the same market with different risk levels α_i.

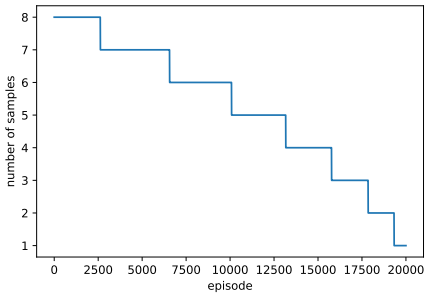

Figure 1. The number of samples of Algorithm 1, 2 and 3.

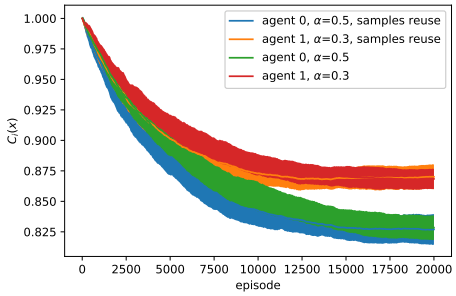

Figure 2. CVaR values achieved by Algorithm 1 (green and red) and Algorithm 2 (blue and orange). The solid lines and shades are averages and standard deviations over 20 runs.

We let $\alpha_0 = 0.5$ and $\alpha_1 = 0.3$, i.e, firm 1 is more risk sensitive than firm 0. Suppose that firm i supplies the market with a quantity x_i, and the total supply of both firms defines the price of the goods in the market. We assume that the cost function for each firm $i \in \{0, 1\}$ is defined as $J_i = -(2-\sum_j x_j)x_i + 0.1x_i + \xi_i x_i + 1$, where $\xi_i \sim U(0, 1)$ is a uniform random variable. The cost term $\xi_i x_i$ models the uncertainty in the market, which is proportional to production. The goal of each firm is to minimize the CVaR of their local cost function.

Recall the definition of the regret in Section 2, and note that it is not possible to compare the performance of Algorithm 1, 2 and 3 in terms of regret, since the baseline term $\min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i(\tilde{x}_i, x_{-i,t})$ depends on the given sequence of $\{x_{-i,t}\}_{t=1}^{T}$ and there is no golden rule to help select this sequence. In addition, computing the analytical solution to the baseline term is challenging due to the variational definition of CVaR. Instead, in what follows, we compare Algorithms 1,2 and 3 in terms of their empirical performances. Specifically, we run Algorithms 1, 2, and 3 and calculate the CVaR values achieved for each action. The algorithm with lower CVaR values is preferred since it achieves better performance in terms of risk-aversion.

The number of samples used by Algorithm 1 is determined by equation (1) and is shown in Figure 1. We implement a hybrid sampling strategy for Algorithm 2 and select the switching time step as 15000, after which prior samples are reused as in equation (14). The reason for this choice is that after 15000 time steps the number of samples becomes small (less than or equal to 3), which causes large errors in CVaR estimation. All other parameters in Algorithms 1, 2 and 3 are tuned so that the three algorithms achieve individually their best performance. Figure 2 compares empirically the performance of Algorithms 1 and 2. We observe that both Algorithms 1 and 2 both converge to the same CVaR values, but Algorithm 2 that reuses samples converges at a faster speed. Indeed, the learning rates of both algorithms depend on the number of samples; Algorithm 2 converges faster because sample reuse increases the effective number of samples per iteration and, as a result, decreases the CVaR estimation errors. This allows for a larger learning rate. Figure 3 shows the variance reduction effect achieved by Algorithm 3 that employs residual feedback to estimate the CVaR gradients. Note the very low variance (almost non-existent) associated with the blue and orange curves. As with Algorithm 2, residual feedback allows Algorithm 3 to use a larger learning rate and still converge to the same CVaR values as Algorithm 1. Additional numerical simulations for different sampling strategies are provided in Appendix D, where we also discuss convergence to the Nash equilibrium in practice.

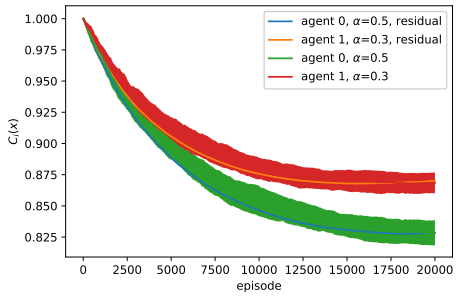

Figure 3. CVaR values achieved by Algorithm 1 (green and red) and Algorithm 3 (blue and orange). The solid lines and shades are averages and standard deviations over 20 runs.

Motivated by the improvements in performance that can be achieved by reusing prior samples to estimate the CVaR values (Algorithm 2) and relying on residual feedback to estimate the CVaR gradients (Algorithm 3), it is of interest to analyze the combined effect of these methods for risk-averse learning in online convex games. However, the theoretical analysis of this method is nontrivial and, for this reason, it is left for future research.

6. Conclusion

In this work, we proposed a first no-regret algorithm for risk-averse online convex games. Our algorithm relied on a new sampling strategy to estimate the CVaR values of the agents'

cost functions, and a zeroth-order estimator of the CVaR gradients to update the agents' actions. To further improve the regret bounds achieved by our algorithm, we proposed two novel modifications; one that reuses samples from the previous iteration to better estimate the CVaR values and another that uses residual feedback to reduce the variance of the CVaR gradient estimation. We illustrated our proposed method on an online market example modeled as a Cournot game.

Acknowledgements

This work is supported in part by AFOSR under award #FA9550-19-1-0169 and by NSF under award CNS-1932011.

References

Allaz, B. and Vila, J.-L. Cournot competition, forward markets and efficiency. *Journal of Economic theory*, 59 (1):1–16, 1993.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Bravo, M., Leslie, D. S., and Mertikopoulos, P. Bandit learning in concave n-person games. *arXiv preprint arXiv:1810.01925*, 2018.

Cardoso, A. R. and Xu, H. Risk-averse stochastic convex bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 39–47. PMLR, 2019.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Curi, S., Levy, K., Jegelka, S., Krause, A., et al. Adaptive sampling for stochastic risk-averse learning. *arXiv preprint arXiv:1910.12511*, 2019.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Duvocelle, B., Mertikopoulos, P., Staudigl, M., and Vermeulen, D. Learning in time-varying games. *arXiv preprint arXiv:1809.03066*, 2018.

Gordon, G. J., Greenwald, A., and Marks, C. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pp. 360–367, 2008.

Hazan, E. *Efficient algorithms for online convex optimization and their applications*. Princeton University, 2006.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Kalogerias, D. S. Noisy linear convergence of stochastic gradient descent for cv@r statistical learning under polyak-lojasiewicz conditions. *arXiv preprint arXiv:2012.07785*, 2020.

Kalogerias, D. S. and Powell, W. B. Zeroth-order stochastic compositional algorithms for risk-aware learning. *arXiv preprint arXiv:1912.09484*, 2019.

Lin, T., Zhou, Z., Mertikopoulos, P., and Jordan, M. Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*, pp. 6161–6171. PMLR, 2020.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Rosen, J. B. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*, 32:13624–13633, 2019.

Shalev-Shwartz, S. and Singer, Y. Convex repeated games and fenchel duality. In *NIPS*, volume 6, pp. 1265–1272. Citeseer, 2006.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Sharpe, W. F. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

Shi, Y. and Zhang, B. No-regret learning in cournot games. *arXiv preprint arXiv:1906.06612*, 2019.

Soma, T. and Yoshida, Y. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

Tamkin, A., Keramati, R., Dann, C., and Brunskill, E. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

Tatarenko, T. and Kamgarpour, M. Learning generalized nash equilibria in a class of convex games. *IEEE Transactions on Automatic Control*, 64(4):1426–1439, 2018.

Tatarenko, T. and Kamgarpour, M. Bandit online learning of nash equilibria in monotone games. *arXiv preprint arXiv:2009.04258*, 2020.

Urpí, N. A., Curi, S., and Krause, A. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

Zhang, Y., Zhou, Y., Ji, K., and Zavlanos, M. M. Boosting one-point derivative-free online optimization via residual feedback. *arXiv preprint arXiv:2010.07378*, 2020a.

Zhang, Y., Zhou, Y., Ji, K., and Zavlanos, M. M. Improving the convergence rate of one-point zeroth-order optimization using residual feedback. *arXiv preprint arXiv:2006.10820*, 2020b.

A. Proofs of Key Results Supporting Algorithm 1

A.1 Proof of Lemma 2

Proof. 1. From the convexity of $C_i(x_i, x_{-i})$, we can get that $C_i(\theta p_1+(1-\theta)p_2, x_{-i}) \leq \theta C_i(p_1, x_{-i})+(1-\theta)C_i(p_2, x_{-i})$ for any $p_1, p_2 \in \mathcal{X}_i^\delta$ and $\theta \in [0,1]$, . Here we denote $\mathbb{E}_{w_i\sim\mathbb{B}_i, u_{-i}\sim\mathbb{S}_{-i}}$ as $\mathbb{E}$ due to space limit. Thus we have

$$\begin{aligned}
&C_i^\delta(\theta p_1 + (1-\theta)p_2, x_{-i})\\
=&\mathbb{E}\big[C_i(\theta p_1 + (1-\theta)p_2 + \delta w_i, x_{-i} + \delta u_{-i})\big]\\
=&\mathbb{E}\big[C_i(\theta(p_1 + \delta w_i) + (1-\theta)(p_2 + \delta w_i), x_{-i} + \delta u_{-i})\big]\\
\leq&\mathbb{E}\big[\theta C_i(p_1 + \delta w_i, x_{-i} + \delta u_{-i}) + (1-\theta)C_i(p_2 + \delta w_i, x_{-i} + \delta u_{-i})\big]\\
=&\theta C_i^\delta(p_1 + \delta w_i, x_{-i}) + (1-\theta)C_i^\delta(p_2 + \delta w_i, x_{-i}),
\end{aligned}$$

which completes the proof.
2. According to the definition of C_i^δ function, we have $|C_i^\delta(x) - C_i^\delta(y)| = |\,\mathbb{E}_{w_i\sim\mathbb{B}_i, u_{-i}\sim\mathbb{S}_{-i}}[C_i(x_i + \delta w_i, x_{-i} + \delta u_{-i}) - C_i(y_i + \delta w_i, y_{-i} + \delta u_{-i})]| \leq \mathbb{E}_{w_i\sim\mathbb{B}_i, u_{-i}\sim\mathbb{S}_{-i}}[L_0 \|x-y\|] \leq L_0 \|x-y\|$. The proof is complete.
3. Since the C_i function is L_0-Lipschitz continuous, we have that

$$\begin{aligned}
|C_i^\delta(x) - C_i(x)| =&|\mathbb{E}[C_i(x_i + \delta w_i, x_{-i} + \delta u_{-i})] - C_i(x)|\\
\leq&L_0 \|(\delta w_i, \delta u_{-i})\|\\
\leq& L_0\delta\sqrt{N},
\end{aligned}$$

where the last inequality is due to the fact that $\|w_i\| \leq 1$, $\|u_i\| \leq 1$. □

A.2 Proof of Lemma 3

Before the proof of Lemma 3, we first give the following result.

Lemma 9. *Let F be the CDF of a non-negative random variable bounded by U, then we have that*

$$\mathbb{E}_F[X-\nu]_+ = \int_0^U (1-F(y))dy - \nu + \int_0^\nu F(y)dy.$$

Proof. It follows that

$$\begin{aligned}
\mathbb{E}_F[X-\nu]_+ &= \mathbb{E}_F[(X-\nu)\mathbf{1}\{X>\nu\}]\\
&= \mathbb{E}_F[(X-\nu)(1-\mathbf{1}\{X\leq\nu\})]\\
&= \mathbb{E}_F[X] - \nu + \mathbb{E}_F[X\mathbf{1}\{X\leq\nu\}] + \nu F(\nu).
\end{aligned}$$

Since $a = \int_0^a dy = \int_0^\infty \mathbf{1}\{y\leq a\}dy$, we obtain that

$$\begin{aligned}
\mathbb{E}_F[X-\nu]_+ &= \mathbb{E}_F[X] - \nu + \nu F(\nu) - \mathbb{E}_F\left[\mathbf{1}\{X\leq\nu\}\int_0^\infty \mathbf{1}\{y\leq X\}dy\right]\\
&= \mathbb{E}_F[X] - \nu + \nu F(\nu) - \int_0^\infty \mathbb{P}_F(y\leq X\leq\nu)dy\\
&= \mathbb{E}_F[X] - \nu + \nu F(\nu) - \int_0^\nu (F(\nu)-F(y))dy\\
&= \int_0^U (1-F(y))dy - \nu + \int_0^\nu F(y)dy.
\end{aligned}$$

This ends the proof. □

Now we are ready to give the proof of Lemma 3.

Proof. According to the CVaR property, we have $\mathrm{CVaR}_\alpha[F] = \nu_F + \frac{1}{\alpha}\mathbb{E}_F[X-\nu_F]_+$, $\mathrm{CVaR}_\alpha[G] = \nu_G + \frac{1}{\alpha}\mathbb{E}_G[X-\nu_G]_+$. Since ν_F is the value that minimizes the $\mathrm{CVaR}_\alpha[F]$, we further obtain that

$$\mathrm{CVaR}_\alpha[F] - \mathrm{CVaR}_\alpha[G] \leq \frac{1}{\alpha}\left(\mathbb{E}_F[X-\nu_G]_+ - \mathbb{E}_G[X-\nu_G]_+\right). \tag{17}$$

Using Lemma 9, we can obtain

$$\begin{aligned}
&\mathrm{CVaR}_\alpha[F] - \mathrm{CVaR}_\alpha[G] \\
\leq &\frac{1}{\alpha}\Big(\int_0^U (1-F(y))dy - \nu_G + \int_0^{\nu_G} F(y)dy - \int_0^U (1-G(y))dy + \nu_G - \int_0^{\nu_G} G(y)dy\Big) \\
\leq &\frac{1}{\alpha}\left(\int_0^U (1-F(y))dy - \int_0^U (1-G(y))dy + \int_0^{\nu_G} (F(y)-G(y))dy\right) \\
\leq &\frac{1}{\alpha}\left(\int_0^U (G(y)-F(y))dy + \int_0^{\nu_G} (F(y)-G(y))dy\right) \\
\leq &\frac{1}{\alpha}\int_{\nu_G}^U (G(y)-F(y))dy \\
\leq &\frac{U}{\alpha}\sup_y |F(y)-G(y)|.
\end{aligned}$$

By symmetry, we can bound $\mathrm{CVaR}_\alpha[G] - \mathrm{CVaR}_\alpha[F]$ as well and the proof is omitted. The proof is complete. □

A.3 Proof of Lemma 4

We first give the DKW inequality, which is helpful in our subsequent analysis.

Lemma 10 (DKW inequality)**.** *Let F be the CDF of a random variable and $\hat{F}$ be the empirical CDF obtained by n i.i.d. samples. For a gien constant $\epsilon > 0$, we have*

$$\mathbb{P}\left\{\sup_y |F(y) - \hat{F}(y)| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

Now we are ready to show the proof of Lemma 4.

Proof. According to the DKW inequality, we have that

$$\mathbb{P}\left\{\sup_y |F_{i,t}(y) - \hat{F}_{i,t}(y)| \geq \sqrt{\frac{\ln(2/\bar{\gamma})}{2n_t}}\right\} \leq \bar{\gamma}. \tag{18}$$

Define the events in (18) as A_t. Recall that $\gamma = \bar{\gamma}T$, then the following holds

$$\sup_y |F_{i,t}(y) - \hat{F}_{i,t}(y)| \leq \sqrt{\frac{\ln(2T/\gamma)}{2n_t}}, \forall t = 1, \ldots, T, \tag{19}$$

with probability at least $1-\gamma$ due to the fact that $1 - \mathbb{P}\{\bigcup_{t=1}^T A_t\} \geq 1 - \sum_{t=1}^T \mathbb{P}\{A_t\} \geq 1 - T\frac{\gamma}{T} \geq 1-\gamma$.

Combining with the sampling strategy defined in equation (1), and applying Lemma 3, the accumulated error of CVaR

estimation can be bounded as follows

$$\begin{aligned}\sum_{t=1}^{T}|\hat{\varepsilon}_{i,t}| &\le \sum_{t=1}^{T}\frac{U}{\alpha_i}\sqrt{\frac{\ln(2T/\gamma)}{2n_t}} \le \sum_{t=1}^{T}\frac{U}{\alpha_i}\sqrt{\frac{\ln(2T/\gamma)}{2(bU^2(T-t+1)^a)}}\\ &= \sum_{t=1}^{T}\frac{1}{\alpha_i}\sqrt{\frac{\ln(2T/\gamma)}{2bt^a}} \le \frac{1}{\alpha_i}\sqrt{\frac{\ln(2T/\gamma)}{2b}}\left(1+\int_{t=1}^{T}\frac{1}{\sqrt{t}}dt\right)\\ &\le \frac{1}{\alpha_i}\sqrt{\frac{\ln(2T/\gamma)}{2b}}\left(1+\frac{1}{1-\frac{a}{2}}(T^{1-\frac{a}{2}}-1)\right) \le \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{b}}T^{1-\frac{a}{2}},\end{aligned}$$

where the last inequality holds due to the fact that $\frac{1}{1-\frac{a}{2}} < 2$, which completes the proof. □

A.4 Proof of Lemma 5

Proof. We first present an observation.
Observation 1: There exist a constant $\Omega > 0$ such that

$$\min_{\tilde{x}_i\in\mathcal{X}_i^{\delta}}\sum_{t=1}^{T}C_i^{\delta}(\tilde{x}_i,x_{-i,t}) \le \min_{\tilde{x}_i\in\mathcal{X}_i}\sum_{t=1}^{T}C_i^{\delta}(\tilde{x}_i,x_{-i,t}) + \Omega L_0\delta T.$$

Proof. Let $x_i^1 = \arg\min_{\tilde{x}_i\in\mathcal{X}_i^{\delta}}\sum_{t=1}^{T}C_i^{\delta}(\tilde{x}_i,x_{-i,t})$, $x_i^2 = \arg\min_{\tilde{x}_i\in\mathcal{X}_i}\sum_{t=1}^{T}C_i^{\delta}(\tilde{x}_i,x_{-i,t})$. According to Lemma 3 in (Tatarenko & Kamgarpour, 2020), $\left\|x_i^1 - x_i^2\right\| = O(\delta)$, i.e, there exist a constant $\Omega > 0$ such that $\left\|x_i^1 - x_i^2\right\| \le \Omega\delta$. Adding that $C_i^{\delta}(x_i,x_{-i})$ is L_0-Lipschitz continuous, we can easily obtain the claim. □

Let $x_{\delta_i}^* = \min_{\tilde{x}_i\in\mathcal{X}_i}\sum_{t=1}^{T}C_i(\tilde{x}_i,x_{-i,t})$. We have that

$$\begin{aligned}\left\|x_{i,t+1}-x_{\delta_i}^*\right\|^2 &= \left\|\mathcal{P}_{\mathcal{X}_i^{\delta}}(x_{i,t}-\eta\hat{g}_{i,t})-x_{\delta_i}^*\right\|^2 \le \left\|x_{i,t}-\eta\hat{g}_{i,t}-x_{\delta_i}^*\right\|^2\\ &= \left\|x_{i,t}-x_{\delta_i}^*\right\|^2 + \eta^2\left\|\hat{g}_{i,t}\right\|^2 - 2\eta\langle\hat{g}_{i,t},x_{i,t}-x_{\delta_i}^*\rangle. \end{aligned} \tag{20}$$

Since $C_i^{\delta}(x_i,x_{-i})$ is convex in x_i, we have that

$$\begin{aligned}&C_i^{\delta}(x_t)-C_i^{\delta}(x_{\delta_i}^*,x_{-i,t}) \le \nabla_i C_i^{\delta}(x_t)(x_{i,t}-x_{\delta_i}^*) = \mathbb{E}\langle\hat{g}_{i,t},x_{i,t}-x_{\delta_i}^*\rangle - \mathbb{E}\langle\frac{d_i}{\delta}\hat{\varepsilon}_{i,t}u_{i,t},x_{i,t}-x_{\delta_i}^*\rangle\\ &\le \frac{1}{2\eta}\mathbb{E}\big(\left\|x_{i,t}-x_{\delta_i}^*\right\|^2 - \left\|x_{i,t+1}-x_{\delta_i}^*\right\|^2\big) + \frac{\eta}{2}\mathbb{E}\left\|\hat{g}_{i,t}\right\|^2 + \mathbb{E}\big[\frac{d_i}{\delta}\left\|\hat{\varepsilon}_{i,t}\right\|\left\|x_{i,t}-x_{\delta_i}^*\right\|\big], \end{aligned} \tag{21}$$

where the last inequality follows from equation (20).

Taking the sum from $t = 1$ to T on both sides of equation (21), we obtain that

$$\begin{aligned}&\sum_{t=1}^{T}C_i^{\delta}(x_{i,t},x_{-i,t}) - \min_{\tilde{x}_i\in\mathcal{X}_i^{\delta}}\sum_{t=1}^{T}C_i^{\delta}(\tilde{x}_i,x_{-i,t})\\ \le&\frac{\left\|x_{i,1}-x_{\delta_i}^*\right\|^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\big[\sum_{t=1}^{T}\left\|\hat{g}_{i,t}\right\|^2\big] + \mathbb{E}\big[\sum_{t=1}^{T}\frac{d_i}{\delta}\left\|\hat{\varepsilon}_{i,t}\right\|\left\|x_{i,t}-x_{\delta_i}^*\right\|\big]\\ \le&\frac{D_x^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\big[\sum_{t=1}^{T}\left\|\hat{g}_{i,t}\right\|^2\big] + \frac{d_iD_x}{\delta}\mathbb{E}\big[\sum_{t=1}^{T}\left\|\hat{\varepsilon}_{i,t}\right\|\big]. \end{aligned} \tag{22}$$

Algorithm 2 Risk-averse learning with sample reuse

Require: Initial value x_0, step size η, parameters a, b, δ, T, risk level α_i, $i = 1, \cdots, N$.

for *episode* $t = 1, \ldots, T$ **do**
  Select $n_t = \lceil bU^2(T-t+1)^a \rceil$
  Each agent samples $u_{i,t} \in \mathbb{S}^{d_i}$, $i = 1, \ldots, N$
  Each agent play $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$, $i = 1, \ldots, N$
  for $j = 1, \ldots, n_t$ **do**
    Let all agents play $\hat{x}_{i,t}$
    Obtain $J_i(\hat{x}_{i,t}, \hat{x}_{-i,t}, \xi_i^j)$
  end for
  for agent $i = 1, \ldots, N$ **do**
    Build EDF $\tilde{F}_{i,t}(y)$
    Calculate CVaR estimate: $\mathrm{CVaR}_{\alpha_i}[\tilde{F}_{i,t}(y)]$
    Construct gradient estimate $\tilde{g}_{i,t} = \frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\tilde{F}_{i,t}(y)]u_{i,t}$
    Update x: $x_{i,t+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta\tilde{g}_{i,t})$
  end for
end for

Recalling that $|C_i^\delta(x) - C_i(x)| \le L_0\sqrt{N}\delta$, and $C_i^\delta(x)$ is Lipschitz continuous, it follows that

$$\begin{aligned}
\mathrm{R}_i^C(T) &= \sum_{t=1}^{T} C_i(\hat{x}_t) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i(\tilde{x}_i, \hat{x}_{-i,t}) \\
&\le \sum_{t=1}^{T} C_i^\delta(\hat{x}_t) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i^\delta(\tilde{x}_i, \hat{x}_{-i,t}) + 2\delta L_0\sqrt{N}T \\
&\le \sum_{t=1}^{T} C_i^\delta(x_t) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i^\delta(\tilde{x}_i, x_{-i,t}) + 4\delta L_0\sqrt{N}T.
\end{aligned}$$

Applying Observation 1 and (22) into the inequality above, it gives that

$$\begin{aligned}
\mathrm{R}_i^C(T) &\le \sum_{t=1}^{T} C_i^\delta(x_t) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^{T} C_i^\delta(\tilde{x}_i, x_{-i,t}) + 4\delta L_0\sqrt{N}T \\
&\le \sum_{t=1}^{T} C_i^\delta(x_t) - \min_{\tilde{x}_i \in \mathcal{X}_i^\delta} \sum_{t=1}^{T} C_i^\delta(\tilde{x}_i, x_{-i,t}) + \Omega L_0\delta T + 4\delta L_0\sqrt{N}T \\
&\le \frac{D_x^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\big[\sum_{t=1}^{T}\|\hat{g}_{i,t}\|^2\big] + \frac{d_i D_x}{\delta}\mathbb{E}\big[\sum_{t=1}^{T}\|\hat{\varepsilon}_{i,t}\|\big] + 4\delta L_0\sqrt{N}T + \Omega L_0\delta T \\
&\le \frac{D_x^2}{2\eta} + \frac{d_i^2U^2\eta}{2\delta^2}T + \frac{d_i D_x}{\delta}\mathbb{E}\big[\sum_{t=1}^{T}\|\hat{\varepsilon}_{i,t}\|\big] + (4\sqrt{N} + \Omega)L_0\delta T, \qquad (23)
\end{aligned}$$

where the last inequality follows from the fact that $|\hat{C}_i| \le U$ and $\|\hat{g}_{i,t}\| = \left\|\frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]u_{i,t}\right\| \le \frac{d_i U}{\delta}$. □

B. Proofs of Key Results Supporting Algorithm 2

B.1 Proof of Lemma 6

Proof. In order to give the formal proof, we need to introduce some new definitions. Define a new random variable $\check{J}_{i,t} = z_t J_{i,t} + (1 - z_t)J_{i,t-1}$, where $J_{i,t}$, $J_{i,t-1}$ are abbreviation of $J_i(\hat{x}_t, \xi_i)$, $J_i(\hat{x}_{t-1}, \xi_i)$, z_t is an independent Bernoulli random variable with $\mathbb{P}\{z_t = 1\} = \frac{n_t}{N_t}$. Define the events $B_t := \{\#(z_t = 1) = n_t \text{ from } N_t \text{ samples of } \check{J}_{i,t}\}$, where $\#$

denotes the times that the event occurs. Then we define the conditional expectation of $\check{J}_{i,t}$ given event B_t as $\bar{J}_{i,t} = \mathbb{E}[\check{J}_{i,t}|B_t]$, which is still a random variable (Durrett, 2019). Define the CDF of $\bar{J}_{i,t}$ as $\bar{F}_{i,t}$, then we have that

$$\begin{aligned}\bar{F}_{i,t}(y) &= \mathbb{P}\{\bar{J}_{i,t} \le y\} = \mathbb{P}\{z_t J_{i,t} + (1-z_t)J_{i,t-1} \le y|B_t\} \\ &= \mathbb{P}\{z_t = 1|B_t\}\mathbb{P}\{J_{i,t} \le y\} + \mathbb{P}\{z_t = 0|B_t\}\mathbb{P}\{J_{i,t-1} \le y\} = \frac{n_t}{N_t}F_{i,t} + \frac{n_{t-1}}{N_t}F_{i,t-1},\end{aligned} \tag{24}$$

where the last equality is obtained by definitions of the random variables z_t, $J_{i,t}$, $J_{i,t-1}$. The second last equality is due to the fact that B_t is only related to z_t, and z_t is independent of $J_{i,t}$,$J_{i,t-1}$. Then it gives that

$$\begin{aligned}&\sup_y |\bar{F}_{i,t}(y) - F_{i,t}(y)| \\ \le &\sup_y |\frac{n_{t-1}}{N_t}(F_{i,t-1}(y) - F_{i,t}(y))| \\ \le &\frac{1}{2}\sup_y |F_{i,t-1}(y) - F_{i,t}(y)|,\end{aligned} \tag{25}$$

where the last inequality is due to $n_{t-1} \le n_t$ and thus $\frac{n_{t-1}}{N_t} \le \frac{1}{2}$. From the definition of $\bar{J}_{i,t}$, we have that $\tilde{F}_{i,t}$ is an EDF of $\bar{F}_{i,t}$. Applying DKW inequality, we have that, for $t \ge 2$

$$\mathbb{P}\left\{\sup_y |\tilde{F}_{i,t}(y) - \bar{F}_{i,t}(y)| \ge \sqrt{\frac{\ln(2/\bar{\gamma})}{2(n_t + n_{t-1})}}\right\} \le \bar{\gamma}. \tag{26}$$

Define the event in (26) as E_t. Then the following holds

$$\sup_y |\tilde{F}_{i,t}(y) - \bar{F}_{i,t}(y)| \le \sqrt{\frac{\ln(2T/\gamma)}{2(n_t + n_{t-1})}}, \forall t = 2, \dots, T, \tag{27}$$

with the probability at least $1-\gamma$ due to the fact that $1 - \mathbb{P}\{\bigcup_{t=2}^T E_t\} \ge 1 - \sum_{t=2}^T \mathbb{P}\{E_t\} \ge 1 - (T-1)\frac{\gamma}{T} \ge 1 - \gamma$. Together with Assumption 3, the following holds

$$\begin{aligned}\sup_y |\tilde{F}_{i,t}(y) - F_{i,t}(y)| =& \sup_y |\tilde{F}_{i,t}(y) - \bar{F}_{i,t}(y) + \bar{F}_{i,t}(y) - F_{i,t}(y)| \\ \le& \sup_y |\tilde{F}_{i,t}(y) - \bar{F}_{i,t}(y)| + \sup_y |\bar{F}_{i,t}(y) - F_{i,t}(y)| \\ \le& \sqrt{\frac{\ln(2/\bar{\gamma})}{2(n_t + n_{t-1})}} + \frac{1}{2}\sup_y |F_{i,t-1}(y) - F_{i,t}(y)| \\ \le& \sqrt{\frac{\ln(2/\bar{\gamma})}{2(n_t + n_{t-1})}} + \frac{C_1\delta + C_2}{2}\left\|x_t - x_{t-1}\right\|,\end{aligned} \tag{28}$$

with probability at least $1-\gamma$ for $\forall t = 2, \dots, T$. Then, applying Lemma 3, it gives that

$$\begin{aligned}|\tilde{\varepsilon}_{i,t}| &\le \frac{U}{\alpha_i}\sup_y |\tilde{F}_{i,t}(y) - F_{i,t}(y)| \\ &\le \frac{U}{\alpha_i}\left(\sqrt{\frac{\ln(2/\bar{\gamma})}{2(n_t + n_{t-1})}} + \frac{C_1\delta + C_2}{2}\left\|x_t - x_{t-1}\right\|\right) \\ &\le \frac{U}{\alpha_i}\left(\sqrt{\frac{\ln(2T/\gamma)}{2(n_t + n_{t-1})}} + \frac{(C_1\delta + C_2)d_i U\sqrt{N}\eta}{2\delta}\right),\end{aligned} \tag{29}$$

where the inequality holds due to the fact that $\|\tilde{g}_{i,t}\| = \left\|\frac{d_i}{\delta}\mathrm{CVaR}_{\alpha_i}[\tilde{F}_{i,t}]u_{i,t}\right\| \le \frac{d_i}{\delta}U$ and $\|x_t - x_{t-1}\| \le \eta\|\tilde{g}_t\| \le \frac{\eta d_i U\sqrt{N}}{\delta}$. The proof is complete. □

B.2 Proof of Lemma 8

Proof. The proof is very similar to the proof of Lemma 5. By substituting $\hat{g}_{i,t}$, $\hat{\varepsilon}_{i,t}$ with $\tilde{g}_{i,t}$, $\tilde{\varepsilon}_{i,t}$, we can obtain the claim. The detailed proof is omitted. □

B.3 Proof of Corollary 1

Proof. According to the sampling strategy in (1), we have that

$$\begin{aligned}\sum_{t=1}^{T}|\tilde{\varepsilon}_{i,t}| &= \sum_{t=1}^{t_0}|\tilde{\varepsilon}_{i,t}| + \sum_{t=t_0+1}^{T}|\tilde{\varepsilon}_{i,t}| \\ &\le \sum_{t=1}^{t_0}|\tilde{\varepsilon}_{i,t}| + \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} := \bar{B}_1(t_0).\end{aligned}$$

When $t \le t_0$, since $\tilde{F}_{i,t} = \hat{F}_{i,t}$, we have that $\tilde{\varepsilon}_{i,t} = \hat{\varepsilon}_{i,t}$. Next we focus on the term $\sum_{t=t_0+1}^{T}|\tilde{\varepsilon}_{i,t}|$. Applying the inequality in (11) and substituting $\delta = \frac{\sqrt{D_x U d_i}}{N^{\frac{1}{4}}T^{\frac{a}{4}}\sqrt{\alpha_i L_0}}$, $\eta = \frac{\sqrt{\alpha_i}D_x^{\frac{3}{2}}}{\sqrt{L_0 U d_i}N^{\frac{1}{4}}T^{\frac{3a}{4}}}$ into it, we have that

$$\begin{aligned}\sum_{t=t_0+1}^{T}|\tilde{\varepsilon}_{i,t}| &\le \frac{U}{\alpha_i}\sum_{t=t_0+1}^{T}\sqrt{\frac{\ln(2T/\gamma)}{4n_{t-1}}} + \frac{(C_1\delta + C_2)d_i U^2\sqrt{N}\eta(T-t_0)}{2\alpha_i\delta} \\ &\le \sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} + \Sigma_5(T-t_0)T^{-\frac{3a}{4}} + \Sigma_6(T-t_0)T^{-\frac{a}{2}},\end{aligned}$$

where $\Sigma_5 = \frac{C_1 D_x^{\frac{3}{2}}N^{\frac{1}{2}}d_i^{\frac{1}{2}}U^{\frac{3}{2}}}{2\alpha_i^{\frac{1}{2}}L_0^{\frac{1}{2}}}$, $\Sigma_6 = \frac{C_2 U N^{\frac{1}{2}}D_x}{2}$.

Set $p = \frac{\ln(T-t_0)}{\ln T}$, then we have $T - t_0 = T^p$ with $p \in (0,1)$. Then, it gives that

$$\begin{aligned}&\sum_{t=1}^{T}|\tilde{\varepsilon}_{i,t}| - \bar{B}_1(t_0) \\ &\le \sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} + \Sigma_5(T-t_0)T^{-\frac{3a}{4}} + \Sigma_6(T-t_0)T^{-\frac{a}{2}} - \sqrt{\frac{2\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} \\ &\le (1-\sqrt{2})\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}T^{p(1-\frac{a}{2})} + \Sigma_5 T^{p-\frac{3a}{4}} + \Sigma_6 T^{p-\frac{a}{2}} \\ &\le T^{p-\frac{a}{2}}\left((1-\sqrt{2})\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}T^{\frac{a}{2}(1-y)} + \frac{\Sigma_5}{T^{\frac{a}{4}}} + \Sigma_6\right),\end{aligned} \tag{30}$$

Define the function $g(T) = (1-\sqrt{2})\sqrt{\frac{\ln(2T/\gamma)}{\alpha_i^2 b}}T^{\frac{a}{2}(1-y)} + \frac{\Sigma_5}{T^{\frac{a}{4}}} + \Sigma_6$. It can be verified that the function $g(T)$ is monotonically deceasing in T and approaches negative infinity when $T \to \infty$. Then there must exist $T_{\lambda(t_0)}$ such that when $T > T_{\lambda(t_0)}$, we have $g(T) \le -\lambda$, and thus $\sum_{t=1}^{T}|\tilde{\varepsilon}_{i,t}| - \bar{B}_1(t_0) \le -\lambda T^{p-\frac{a}{2}}$. Recall that $\bar{B}_1(t_0) = \sum_{t=1}^{t_0}|\tilde{\varepsilon}_{i,t}| + \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} \le \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{\alpha_i^2 b}}(T^{1-\frac{a}{2}} - (T-t_0)^{1-\frac{a}{2}}) + \frac{1}{\alpha_i}\sqrt{\frac{2\ln(2T/\gamma)}{\alpha_i^2 b}}(T-t_0)^{1-\frac{a}{2}} = B_1$. By virtue of Lemma 8, it gives that

$$\begin{aligned}\mathrm{R}_{C_i}^2(T) &\le Err(ZO) + \frac{d_i D_x}{\delta}\sum_{t=1}^{T}|\tilde{\varepsilon}_{i,t}| \\ &\le Err(ZO) + Err(CVaR) + \frac{d_i D_x}{\delta}(\sum_{t=1}^{T}|\tilde{\varepsilon}_{i,t}| - \bar{B}_1(t_0)) \\ &\le Err(ZO) + Err(CVaR) - \lambda T^{p-\frac{a}{2}}.\end{aligned}$$

By choosing $\delta = \frac{\sqrt{D_x U d_i}}{N^{\frac{1}{4}} T^{\frac{a}{4}} \sqrt{\alpha_i L_0}}$, $\eta = \frac{\sqrt{\alpha_i} D_x^{\frac{3}{2}}}{\sqrt{L_0 U d_i} N^{\frac{1}{4}} T^{\frac{3a}{4}}}$, the result follows from the same proof as in Theorem 1. The proof is complete. □

C. Proofs of Key Results Supporting Algorithm 3

Algorithm 3 Risk-averse learning with residual feedback

Require: Initial value x_0, step size η, parameters a, b, δ, T, risk level α_i, $i = 1, \cdots, N$.

for *episode* $t = 1, \ldots, T$ **do**
  Select $n_t = \lceil bU^2(T-t+1)^a \rceil$
  Each agent samples $u_{i,t} \in \mathbb{S}^{d_i}$, $i = 1, \ldots, N$
  Each agent play $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$, $i = 1, \ldots, N$
  for $j = 1, \ldots, n_t$ **do**
    Let all agents play $\hat{x}_{i,t}$
    Obtain $J_i(\hat{x}_{i,t}, \hat{x}_{-i,t}, \xi_i^j)$
  end for
  for agent $i = 1, \ldots, N$ **do**
    Build EDF $\hat{F}_{i,t}(y)$
    Calculate CVaR estimate: $\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]$
    Construct gradient estimate $\bar{g}_{i,t} = \frac{d_i}{\delta}\left(\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]\right) u_{i,t}$
    Update x: $x_{i,t+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta \bar{g}_{i,t})$
  end for
end for

Observe that the zeroth-order CVaR gradient satisfies the following inequality.

Lemma 11.

$$\sum_{t=1}^{T} \|\bar{g}_{i,t}\|^2 \leq \frac{1}{1-\beta}\|\bar{g}_1\|^2 + \frac{16 d_i^2 N^2 L_0^2}{1-\beta} T + \frac{4 d_i^2 \ln(2T/\gamma) S(\alpha)}{(1-\beta) b \delta^2} T, \tag{31}$$

where $\beta = \frac{4 d_i^2 L_0^2 N \eta^2}{\delta^2}$.

C.1 Proof of Lemma 11

Proof. According to the definition of $\bar{g}_{i,t}$ in (15), it gives that

$$\begin{aligned}
\|\bar{g}_{i,t}\|^2 &= \frac{d_i^2}{\delta^2}\left((\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}])u_{i,t}\right)^2 \\
&\leq \frac{d_i^2}{\delta^2}\left(\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]\right)^2 \|u_{i,t}\|^2 \\
&\leq \frac{d_i^2}{\delta^2}\left(2(\mathrm{CVaR}_{\alpha_i}[F_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t-1}])^2 + 2(\hat{\varepsilon}_{i,t} - \hat{\varepsilon}_{i,t-1})^2\right)\|u_{i,t}\|^2,
\end{aligned} \tag{32}$$

where the last inequality is due to the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. Note that even though $\mathrm{CVaR}_{\alpha_i}[F_{i,t}]$ is Lipschitz continuous, the $\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}]$ is not. We can not bound the difference $\mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\hat{F}_{i,t-1}]$ directly. Then, by virtue of the Lipschitz property of the function $C_i(x)$, we have that

$$\begin{aligned}
&(\mathrm{CVaR}_{\alpha_i}[F_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t-1}])^2 = (C_i(\hat{x}_t) - C_i(\hat{x}_{t-1}))^2 \\
&\leq L_0^2 \|\hat{x}_t - \hat{x}_{t-1}\|^2 \leq L_0^2 \|x_t - x_{t-1} + \delta u_t - \delta u_{t-1}\|^2 \leq L_0^2 (2\|x_t - x_{t-1}\|^2 + 2\|\delta u_t - \delta u_{t-1}\|^2) \\
&\leq 2L_0^2 \|x_t - x_{t-1}\|^2 + 2\delta^2 (2\|u_t\|^2 + 2\|u_{t-1}\|^2) \leq 2L_0^2 \|x_t - x_{t-1}\|^2 + 8L_0^2 N \delta^2,
\end{aligned} \tag{33}$$

where the last inequality is because $\|u_t\|^2 = \sum_{j=1}^N \|u_{i,t}\|^2 = N$. Recall that $x_{i,t} = \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t-1} - \eta_i \bar{g}_{i,t-1})$, we get that $\|x_{i,t} - x_{i,t-1}\| = \left\|\mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t-1} - \eta \bar{g}_{i,t-1}) - \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t-1})\right\| \leq \eta \|\bar{g}_{i,t-1}\|$. Substituting this into inequality (33), we have

that

$$(\mathrm{CVaR}_{\alpha_i}[F_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t-1}])^2 \leq 2L_0^2\eta^2 \left\|\bar{g}_{t-1}\right\|^2 + 8L_0^2 N\delta^2,$$

where $\bar{g}_{t-1}$ is a concatenated vector of $\bar{g}_{i,t-1}$, $i = 1\ldots, N$. According to Lemma 4, with probability $1-\gamma$, we have$|\hat{\varepsilon}_{i,t}|^2 \leq \frac{U^2}{\alpha_i^2}\frac{\ln(2T/\gamma)}{2n_t} \leq \frac{U^2}{\alpha_i^2}\frac{\ln(2T/\gamma)}{2bU^2(T-t+1)} \leq \frac{\ln(2T/\gamma)}{2b\alpha_i^2}$. Applying this inequality and (33) into (32), it gives that

$$\begin{aligned}\left\|\bar{g}_{i,t}\right\|^2 &\leq \frac{d_i^2}{\delta^2}(4L_0^2\eta^2\left\|\bar{g}_{t-1}\right\|^2 + 16L_0^2N\delta^2 + \frac{4\ln(2T/\gamma)}{b\alpha_i^2}) \\ &\leq \frac{4d_i^2L_0^2\eta^2}{\delta^2}\left\|\bar{g}_{t-1}\right\|^2 + 16d_i^2NL_0^2 + \frac{4d_i^2\ln(2T/\gamma)}{b\alpha_i^2\delta^2}.\end{aligned}$$

Summing up this inequality over $i = 1, \ldots, N$, and setting $\sum_i \frac{1}{\alpha_i^2} = S(\alpha)$, we obtain that

$$\left\|\bar{g}_t\right\|^2 \leq \frac{4d_i^2L_0^2N\eta^2}{\delta^2}\left\|\bar{g}_{t-1}\right\|^2 + 16d_i^2N^2L_0^2 + \frac{4d_i^2\ln(2T/\gamma)S(\alpha)}{b\delta^2}.$$

Note that $\beta = \frac{4d_i^2L_0^2N\eta^2}{\delta^2}$. Telescoping this inequality and rearranging the term, we have that

$$\sum_{t=1}^{T}\left\|\bar{g}_t\right\|^2 \leq \frac{1}{1-\beta}\left\|\bar{g}_1\right\|^2 + \frac{16d_i^2N^2L_0^2}{1-\beta} + \frac{4d_i^2\ln(2T/\gamma)S(\alpha)}{(1-\beta)b\delta^2}T.$$

The proof ends by using $\|\bar{g}_{i,t}\| \leq \|\bar{g}_t\|$. □

C.2 Proof of Theorem 3

Proof. With the result in (16) and following similar lines as in the proof of Lemma 5, it can be verified that the following holds

$$\mathrm{R}^3_{C_i}(T) \leq \frac{D_x^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\big[\sum_{t=1}^{T}\left\|\bar{g}_{i,t}\right\|^2\big] + \frac{d_iD_x}{\delta}\mathbb{E}\big[\sum_{t=1}^{T}\left\|\hat{\varepsilon}_{i,t}\right\|\big] + 4\delta L_0\sqrt{N}T + \Omega L_0\delta T,$$

where $\hat{\varepsilon}_{i,t}$ as defined in (10). Applying Lemmas 4 and 11 and substituting the bounds into the inequality above, it gives that

$$\begin{aligned}\mathrm{R}^3_{C_i}(T) \leq &\frac{D_x^2}{2\eta} + \frac{\left\|\bar{g}_1\right\|^2\eta}{2(1-\beta)} + \frac{8d_i^2N^2L_0^2\eta}{1-\beta} + \frac{2d_i^2\ln(2T/\gamma)S(\alpha)\eta}{(1-\beta)b\delta^2}T \\ &+ \frac{d_iD_x\sqrt{2\ln(2T/\gamma)}}{\alpha_i\delta\sqrt{b}}T^{1-\frac{a}{2}} + 4\delta L_0\sqrt{N}T + \Omega L_0\delta T.\end{aligned}$$

Recall that $\eta = \frac{D_x}{d_iL_0N}T^{-\frac{3a}{4}}$, $\delta = \frac{D_x}{N^{\frac{1}{6}}}T^{-\frac{a}{4}}$, we have $\beta = \frac{4d_i^2L_0^2N\eta^2}{\delta^2} = 4N^{\frac{2}{3}}T^{-a} \leq \frac{1}{2}$ when $T \geq (8N^{\frac{2}{3}})^{\frac{1}{a}}$. Therefore, we have $\frac{1}{1-\beta} \leq 2$, and it follows that

$$\begin{aligned}\mathrm{R}^3_{C_i}(T) \leq &\frac{D_x^2}{2\eta} + \left\|\bar{g}_1\right\|^2\eta + 16d_i^2N^2L_0^2\eta + \frac{4d_i^2\ln(2T/\gamma)S(\alpha)\eta}{b\delta^2}T \\ &+ \frac{d_iD_x\sqrt{2\ln(2T/\gamma)}}{\alpha_i\delta\sqrt{b}}T^{1-\frac{a}{2}} + 4\delta L_0\sqrt{N}T + \Omega L_0\delta T.\end{aligned}$$

Substituting δ and η into this inequality and we can get $\mathrm{R}^3_{C_i}(T) = \mathcal{O}(D_xd_iL_0NS(\alpha)\ln(T/\gamma)T^{1-\frac{a}{4}})$. The proof is complete. □

D. Additional Numerical Experiments

We provide additional simulation results for different sampling strategies related to different confidence levels γ. Two different sampling strategies are shown in Figure 4. Figure 5 shows that Algorithm 1 converges faster if more samples are collected and large sample size can decrease the variance of the algorithm.

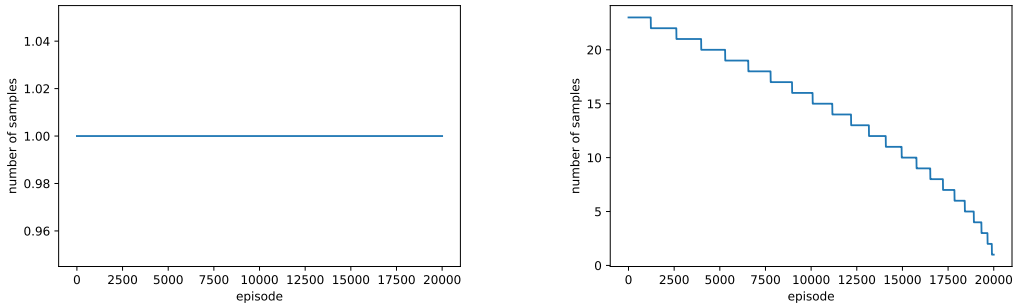

Figure 4. The different choices of number of samples of Algorithm 1.

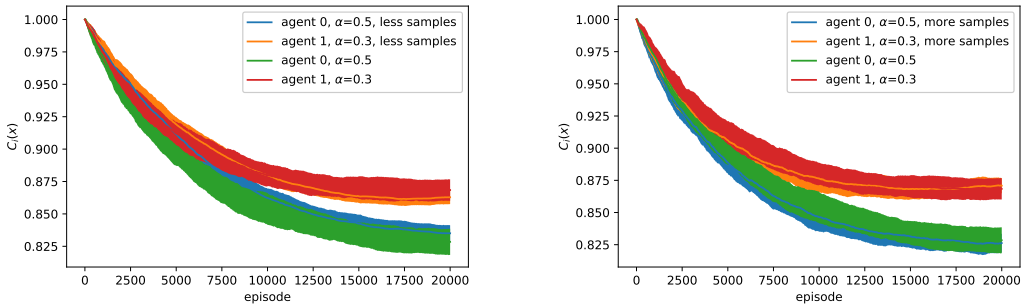

Figure 5. CVaR values achieved by Algorithm 1 with different choices of samples. The solid lines and shades are averages and standard deviations over 20 runs.

Moreover, since risk-neural learning is a special case of risk-averse learning by selecting $\alpha_i = 1$, $i = 1, \ldots, N$, our risk-averse learning algorithms can be also used for risk-neutral games. Although in risk-averse games it is typically hard to calculate the Nash equilibrium, in risk-neutral games calculating the Nash equilibrium is possible. We define the expected cost function of agent i as $u_i = \mathbb{E}_{\xi_i}[J_i] = \mathbb{E}_{\xi_i}[-(2 - \sum_j x_j)x_i + 0.1x_i + \xi_i x_i + 1] = -(2 - \sum_j x_j)x_i + 0.6x_i + 1$, where $\xi_i \sim U(0, 1)$. With this expected cost function, this risk-neural game becomes a convex and monotone game. For this class of games, (Rosen, 1965) has shown that a unique Nash equilibrium exists. Specifically, setting the gradients of the expected cost function equal to 0, we can calculate the Nash equilibrium as $(0.467, 0.467)$. Figure 6 shows that our algorithm converges to a neighborhood of the Nash equilibrium. This experiment verifies the correctness of our methodology.

Figure 6. Action values for risk-neural agents.